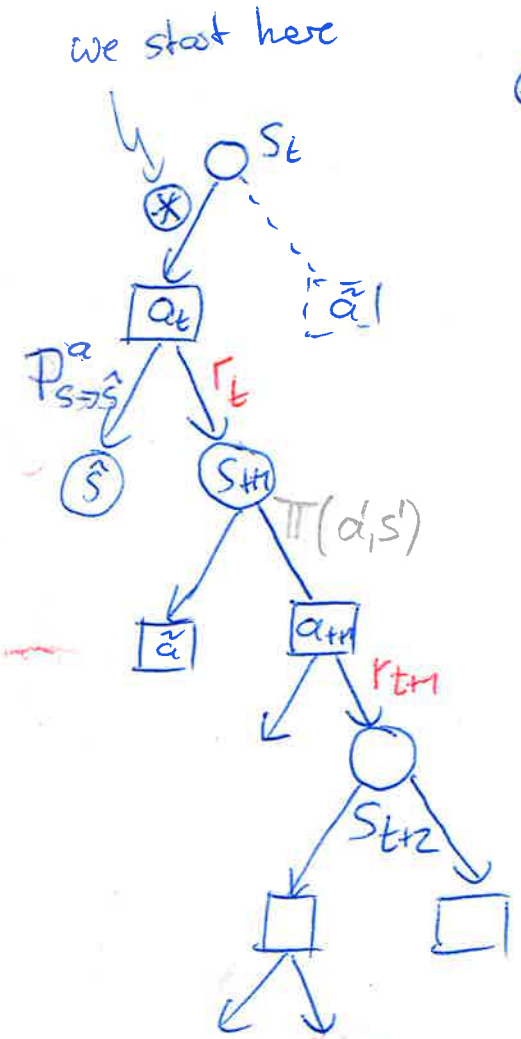
Blackboard RL1-4 Bellman equation



⊛ total reward collected in single trial starting in s with action a_t

$$R^{tot}(s_t,a_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots$$

$$= r_t + \gamma \left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots \right]$$

$$= r_t + \gamma \cdot R^{tot}(s_{t+1}, a_{t+1})$$

total reward (single trial) starting from $s' = s_{t+1}$ with a_{t+1}

now we look at diagram to calculate expectation

$$E\left(R^{tot}(s_t,a_t)\right) = E\left(r_t + \gamma R^{tot}(s_{t+1}, a_{t+1})\right)$$

↑ all future states and actions, given (s_t, a_t)

$$= \sum_{s'} P^{a_t}_{s\to s'} \left[R^{a_t}_{s\to s'} + \gamma E\left(R^{tot} | s'\right) \right]$$

↑ starting in s'

$$= \sum_{s'} P^{a_t}_{s\to s'} \left[R^{a_t}_{s\to s'} + \gamma \cdot \sum_{a'} \pi(a',s') E\left[R(s',a')\right] \right]$$

$$Q(s_t,a_t) = \sum_{s'} P^{a_t}_{s\to s'} \left[R^{a_t}_{s\to s'} + \gamma \sum_{a'} \pi(a',s') \, Q^{tot}(s',a') \right]$$

Blackboard RL-5 — SARSA

from diagram

$$Q(s,a) \approx r_t + \gamma \cdot Q(s',a')$$

(discount: γ)

$$0 \approx r_t + \gamma \cdot Q(s',a') - Q(s,a)$$

proposed update

$$\Delta Q(s,a) = \eta \left[r_t + \gamma \cdot Q(s',a') - Q(s,a) \right]$$

check:

if $r_t > \underbrace{\gamma \, Q(s',a') - Q(s,a)}_{\text{expected reward for this transition}} \Rightarrow$ increase $Q(s,a)$

r_t: actual reward

Blackboard 6A : Consistency of Q-values

assumption:

- set of Q-values is fixed ("frozen")

$$\Leftrightarrow \quad 0 = \mathbb{E}\left[\Delta Q(s,a) \mid s,a\right]$$

$\mathbb{E}$: expectation of next possible states/actions [with frozen policy consistent with frozen Q-values]

$\Rightarrow$ Q-values consistent with Bellman eq.

proof : expectation over possible paths starting in (s,a)

$$0 = \frac{1}{\eta}\,\mathbb{E}\left[\Delta Q(s,a)\right] =$$

$$= \mathbb{E}\left[\frac{1}{\eta}\,\Delta Q(s,a)\right]$$

$$= \mathbb{E}\left[r_t + \gamma\, Q(s',a') - Q(s,a)\right]$$

$$= \mathbb{E}\left[r_t + \gamma\, Q(s',a')\right] - Q(s,a)$$

$Q(s',a')$: frozen

$Q(s,a)$: frozen! trivial expectation

shift $Q(s,a)$ to left-hand side

$$Q(s,a) = \sum_{s'} P^{a}_{s\to s'}\left[R^{a}_{s\to s'} + \sum_{a'} \pi(s',a')\, Q(s',a')\right]$$

$\Rightarrow$ Bellman equation. ∎

Consistency of fluctacting online SARSA with Bellman equation

hypothesis: $0 = \langle \Delta \hat{Q}(s,a) \rangle = \eta \langle r_t + \gamma \hat{Q}(s',a') - \hat{Q}(s,a) \rangle$

cut — shift

$\langle \hat{Q}(s,a) \rangle = \langle r_t + \gamma \cdot \hat{Q}(s',a') \rangle$ — temporal average over many "trials" ($N \to \infty$)

↑ fluctuates — ↑ fluctuates

$$\langle \hat{Q}(s,a) \rangle = \sum_{s'} P^a_{s \to s'} \left[R^a_{s \to s'} + \gamma \sum_{a'} \langle \pi^Q(s',a') \cdot \hat{Q}(s',a') \rangle \right]$$

Problem: π^Q depends on $\hat{Q}$

Solution: ① if η is small, the fluctations of $\hat{Q}$ are small and fluctuations of policy π^Q are "even smaller"*

② consider π^Q fixed for small enough Q

$\Rightarrow$ move π out: $\langle \pi^{\hat{Q}} Q \rangle \rightsquigarrow \pi^Q \langle Q \rangle$

$$\langle \hat{Q}(s,a) \rangle = \sum_{s'} P^a_{s \to s'} \left[R^a_{s \to s'} + \gamma \sum_{a'} \pi^Q(s',a') \langle \hat{Q}(s',a') \rangle \right]$$

$\hat{Q}(s',a') = Q(s',a')$

$\langle \hat{Q}(s,a) \rangle = Q(s,a)$ solves Bellman equation

remarks

① *example of π^Q "even smaller": ε-greedy

only rank-order of Q-values matters: best / 2nd best / ...

if fluctuations $|\Delta \hat{Q}(s',a')| \ll Q(s', \text{best}) - Q(s', 2^{\text{nd}} \text{best})$

then π^Q remains stable!

② evaluation of averages — look at graph

— if in state s' all remaining averages are "given s'"